USER CLASSIFICATION FROM DATA VIA DEEP SEGMENTATION FOR SEMI-SUPERVISED LEARNING

BACKGROUND

[0001] The following relates generally to machine learning, and more specifically, to user classification with semi-supervised machine learning.

[0002] A variety of machine learning techniques may be used to classify data. For example, supervised learning maps inputs to outputs based on observed input-output pairs (i.e., the ground truth). In other words, supervised learning is based on labeled training data. Each piece of labeled training data includes an input object (e.g., a vector) and a desired output value. A supervised learning algorithm learns from the training data to produce a predictive model that may be used to map new input data for which the output is not known.

[0003] By contrast, unsupervised learning draws inferences from datasets consisting of input data without labeled responses. Unsupervised learning may be used to find hidden patterns or grouping in data. For example, cluster analysis is a form of unsupervised learning. Clusters may be identified using measures of similarity such as Euclidean or probabilistic distance.

[0004] In some cases, a dataset may include a large amount of unlabeled data, and a smaller amount of labeled data. In this case, neither supervised nor unsupervised learning can take into account everything that is known about the data. Therefore, there is a need in the art for machine learning techniques that utilize both labeled and unlabeled data.

SUMMARY

[0005] A method, apparatus, and non-transitory computer-readable medium for user classification with semi-supervised machine learning are described. Embodiments of the method, apparatus, and non-transitory computer-readable medium may receive user information for a first set of users, receive survey data for a second set of users, wherein the second set of users is a proper subset of the first set of users, train a first neural network and a second neural network based on the second set of users, wherein the first neural network maps the user information to an embedding space and the second neural network maps the embedding space to a space of probability vectors, and wherein each vector in the space of probability vectors corresponds to a user's category membership propensity, map the user information for the first set of users to the embedding space using the first neural network, predict category membership propensities for the first set of users using a low-density separation algorithm on the user information for the first set of users mapped to the embedding space, update the first neural network and the second neural network based on the prediction, and reclassify the first set of users based on the updated first neural network and the updated second neural network.

[0006] A method, apparatus, and non-transitory computer-readable medium for user classification with semi-supervised machine learning are described. Embodiments of the method, apparatus, and non-transitory computer-readable medium may identify user information for each of a first set of users, select a second set of users, wherein the second set of users is a proper subset of the first set of users, collect additional information for each of the second set of users, segment the second set of users into multiple categories based on the additional information, train a first neural network and a second neural network based on the segmentation of the second set of users, wherein the first neural network maps the user information to an embedding space and the second neural network maps the embedding space to a space of probability vectors corresponding to the multiple categories, map the user information for the first set of users to the embedding space using the first neural network, segment the first set of users into the multiple categories using a low-density separation algorithm, update the first neural network and the second neural network based on the segmentation of the first set of users, and reclassify the first set of users based on the updated first neural network and the updated second neural network.

[0007] A method, apparatus, and non-transitory computer-readable medium for user classification with semi-supervised machine learning are described. Embodiments of the method, apparatus, and non-transitory computer-readable medium may identify information for a first set of objects, identify additional information for a second set of objects, wherein the second set of objects is a subset of the first set of objects, classify the second set of objects into multiple categories based on the additional information, train a first neural network and a second neural network based on the segmentation of the second set of objects, wherein the first neural network maps the object information to an embedding space and the second neural network maps the embedding space to a space corresponding to the multiple categories, map the first set of objects to the embedding space using the first neural network, classify the first set of objects into the multiple categories using a low-density separation algorithm based on the mapping, retrain the first neural network and the second neural network based on the classification of the first set of objects, remap the first set of objects to the embedding space using the updated first neural network, reclassify the first set of objects into the multiple categories using the low-density separation algorithm based on the remapping, and retrain the first neural network and the second neural network based on the reclassification of the first set of objects.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] FIG. **1** shows an example of a comparison of supervised and semi-supervised classification according to aspects of the present disclosure.

[0009] FIG. **2** shows an example of a process for user classification with semi-supervised machine learning according to aspects of the present disclosure.

[0010] FIG. **3** shows an example of a process for receiving user information for a first set of users according to aspects of the present disclosure.

[0011] FIG. **4** shows an example of a process for receiving the survey data according to aspects of the present disclosure.

[0012] FIG. **5** shows an example of a process for training a first neural network and a second neural network according to aspects of the present disclosure.

[0013] FIG. **6** shows an example of separation algorithms according to aspects of the present disclosure.